PhD thesis: Mixed data temporal clustering for modelling longitudinal surveys

Supervision

Prof. Julien JACQUES, ERIC, Université Lyon 2
Prof. Isabelle PRIM-ALLAZ, COACTIS, Université Lyon 2

Context

In many areas of humanities and social sciences, the studies are based on questionnaires completed by participants. Often, these questionnaires are completed several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population. But the statistical analysis of these questionnaires is far from simple, for several reasons. First, the answers to the questions are often of different types: nominal categorical (for example "what is your socio-professional category?"), ordinal categorical (for example "what is your level of satisfaction: bad, average, good?"), quantitative ("what is your age?"), textual (for open questions with free answer). The analysis of such mixed data is a current research problem in the fields of statistics and machine learning, and for lack of an existing solution the practitioner often tends to transform the data to standardize them. Such approach is not satisfying since it leads either to the introduction of a bias or to an important information loss.
The second scientific obstacle is the modelling of the temporal evolution of the answers to the questions. Currently, the analyses are done independently at each temporal phase, then researchers try a posteriori to find links between these different analyses, by seeking from one phase to the other to find similar typical behaviour. The ideal way to model these data would be to propose a model of the temporal evolution, which models all the responses to the questionnaires at the same time. Thus, the analysis will exhibit typical temporal evolution behaviours, which are the objects which researchers in human and social sciences wish to study.

This thesis will thus provide a complete tool for analysing questionnaires repeated over time. The core of the thesis will be the development of a statistical model and associated inference algorithms. But the PhD student will go as far as the implementation of a software tool in the form of an R package, so that researchers in humanities and social sciences can easily use these results.

PhD project

Need of modelling tools for analysing questionnaires repeated over time

In many areas of humanities and social sciences the studies are based on questionnaires completed by participants. The data provided by the answer of participants are of different nature, generally quoted as mixed data in the literature:

- nominal categorical: when questions are of type "what is your socio-professional category?"
- ordinal categorical: when questions are of type "what is your level of satisfaction: bad, average, good?",
- quantitative: when questions are of type "what is your age?",
- textual: for open questions with free answers.

Moreover, we consider repeated questionnaires over time: the participants filled in the questionnaires at several times along the period of study. A time component is then added to the mixed data set.

The machine learning task to which we want answer for this data is unsupervised: there is no specific notions that we want to predict, but we want to explore the data sets in order to exhibit typical behaviours. This task is known as to clustering: we want to build clusters of data such that observations within a cluster are similar and clusters are different from each other. Thus, the data analysis will no longer be based on the observation of the individual responses to the questionnaires, but on the summaries provided by the clusters.
More specifically, once the whole data set of observations over time will be clustered, the clusters will gather set of participants which have the same evolution of answers over time. This information is essential for data analysis from a humanities and social sciences point of view.

State of the art

The classic strategy currently used for the analysis of questionnaires consists of independently analysing the questionnaires of each time phase, and this by standardizing the answers to the questions so that they are all of the same type: either all categorical nominal or all quantitative. Thus, if we consider the example of ordinal data (evaluation on an ordered scale), which are certainly the type most encountered in questionnaires, they are either transformed according to a Likert scale into quantitative data [1,2], or transformed into nominal data by ignoring the order [3]. In the first case, even if there is a whole literature on the construction of Likert scales, the introduction of a notion of distance between categories necessarily brings a bias in the analysis [4]. In the second case, less often used nevertheless, one loses essential information by not taking into account the notion of order within the categories.

Clustering with mixed data have received a large attention in the last decade from the researcher in statistics and machine learning. The latent class model [5] is frequently used. It assumes that the variables are conditionally independent upon the cluster membership. Consequently, the joint probability distribution function (PDF) of the features of different types is obtained by the product of the PDFs of each individual feature (see an implementation using Mixtcomp software [6]). However, when the variables are inherently correlated in a cluster, this model is not suitable. To overcome this issue, the authors of [7] want to conserve standard marginal distributions but also try to loosen the conditional independence on the variables. For this purpose, they use copula, which allow definition of both the dependence model and the type of marginal distributions. The proposed model relies on the main assumption that each cluster follows a Gaussian copula. However, the authors note that model complexity increases with the number of variables, which is not suitable in a big-data context. Another way to address the issues of heterogeneous data is to see some variables as the manifestation of a latent vector. For example, in [8], the clustMD model considers continuous and categorical data (nominal and ordinal) and assumes that a categorical variable is the representation of an underlying latent continuous variable. Then, it is assumed that the continuous variables (observed and unobserved) follow a multivariate Gaussian mixture model. In [9], the authors allow the introduction of more complex data such as functional data or networks by projecting the data set into a reproducing kernel Hilbert space. In [10], another model-based approach for ordinal, nominal, integer and continuous data is proposed, on the basis of conditional independence assumption and with the particularity of creating clusters of features as well as clusters of individuals.

Regarding the temporal dynamics of the data, the main approaches are based on independent analyses. We can for example cite [11] in the case of clustering of ordinal data for an application in psychology. The few existing models for the temporal dynamic focus on a single type of data, as for example [12,13] in the case of quantitative data.

Scientific obstacles and the approaches envisaged to remove them

If [8,9,10] provide several interesting clustering algorithms for mixed data, on challenge remains open: how to specify the weights corresponding to each component of these heterogeneous data in the joint model. In the simple case of joint Gaussian and multinomial data, even after assuming independence leading to a product form, the weight of both types of variable might not always be symmetric. This thesis aims to answer this latter challenge from a modelling point a view, without resorting to some artificially chosen weighting, as is commonly done in the current literature.
The second challenge consists in incorporating the time evolution into the framework. Owing to the fact that the proposed model will enforce a parametric approach, the dynamical component will be introduced via appropriate modelling of the parameter's dynamics through potential smoothness or carefully designed jump processes when necessary. If existing approaches exist in the case of quantitative data [12], all remains to do in the case of more complex data sets.

Knowledge transfer

The results of this work will be published both in leading methodological journals in statistics and also in journals in humanities and social sciences (more precisely business and management).
In addition to the scientific publications, free software (R packages) will be developed in order to be usable for the practitioners in humanities in social sciences. These software tools could also be promoted through specific publications.
This work could in particular be used to shed new light on recent studies carried out by the COACTIS laboratory. In particular for the recent study [14] on changes in eating behaviour during confinement which took place in France in 2020 due to the Covid19 pandemic for which a longitudinal approach has been implemented (4 data collections).

References

[1] Lewis, S.J.G., Foltynie, T., Blackwell, A.D., Robbins, T.W., Owen, A.M. and Barker, R.A. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach, Journal of Neurology, Neurosurgery & Psychiatry, 76 [3], 343--348, 2005.
[2] Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data: An introduction to cluster analysis, John Wiley and Sons Inc., 2008.
[3] Vermunt J.K. and Magidson J. Latent GOLD 4.0 User's Guide. Belmont, Massachusetts, Statistical Innovations Inc., 2005.
[4] Jacques J. and Biernacki C. Model-based co-clustering for ordinal data, Computational Statistics and Data Analysis, 123, 101-115, 2018.
[5] Everitt B.S. Introduction to Latent Variable Models, Chapman and Hall, 1984.
[6] Biernacki C., Deregnaucourt T. and Kubicki V. Model-based clustering with mixed/missing data using the new software MixtComp, CMStatistics 2015 (ERCIM 2015), London, United Kingdom, 2015.
[7] Marbac M., Biernacki C. and Vandewalle V. Model-based clustering of gaussian copulas for mixed data, Communications in Statistics - Theory and Methods, 46 (23), 2017.
[8] McParland D. and Gormley I. Model based clustering for mixed data: Clustmd, Advances in Data Analysis and Classification,10 (2), 155-169, 2016.
[9] Bouveyron C., Fauvel M., Girard S., Kernel discriminant analysis and clustering with parsimonious gaussian process models, Statistics and Computing, 25 (6), 1143-1162, 2015.
[10] Selosse M., Jacques J., Biernacki C. Model-based co-clustering for mixed type data, Computational Statistics and Data Analysis,144, 2020.
[11] Selosse M., Jacques J., Biernacki C. and Cousson-Gélie F. Analyzing health quality survey using constrained co-clustering model for ordinal data and some dynamic implication, Journal of the Royal Statistical Society, Series C, 68 [5], 1327-1349, 2019.

[12] Hasnat M.D.A., Velcin J., Bonnevay S., and Jacques J. Evolutionary clustering for categorical data using parametric links among multinomial mixture models. Econometrics and Statistic, 3, 141-159, 2017.
[13] Widiputra H., Kho H., Lukas, Pears R. and Kasabov N. A Novel Evolving Clustering Algorithm with Polynomial Regression for Chaotic Time-Series Prediction, ICONIP 2009:
Neural Information Processing, 114-121, 2009.
[14] François-Lecompte A., Innocent M., Kréziak D., Prim-Allaz I. Confinement et comportements alimentaires : Quelles évolutions en matière d'alimentation durable ? Revue Française de Gestion, 46, 293, 2020 (à paraître).
[15] Bouveyron C., Celeux G., Murphy B. and Raftery A., Model-based Clustering and Classification for Data Science, with Applications in R, in Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.